
# Characteristics of Verb Use in a Corpus of Child-Directed Japanese

*Tsukuba Psychological Research*
2000, **22**, 29-43



# Characteristics of Verb Use in a Corpus of Child-Directed Japanese

Eleanor Olds Batchelder (*Institute of Psychology, University of Tsukuba, Tsukuba 305-8572, Japan*)

In order to gather quantitative data about Japanese verb inflection, a corpus of child-directed language was studied. In particular, past tense, negative, and *-te*-type forms (*-te, -tara, -tari*) were counted, and these were further categorized according to inflectional paradigm —1-dan, 5-dan, or *su/suru*-type. It was found that, of the three inflection types, the least often used in this corpus were negatives, and the most used were past-tense forms. In addition, by token counts, 5-dan verb forms were found to be most frequent in the combined groups, but in the negative group there were more 1-dan than others. These findings will be used to help interpret the results of a psycholinguistic experiment which asked subjects to provide such inflected forms for made-up verbs.

**Key words**: verb, Japanese, inflection, frequency, types and tokens.

Psycholinguistic researchers have suggested a number of factors which are thought to influence the cognitive processing of language, including frequency in the language, semantic and phonological similarity, imageability, and so forth. In order to design and interpret psycholinguistic experiments, some estimate of these characteristics for each linguistic stimulus is necessary. They may be determined by introspection, by surveys of language users, or by quantitative statistics taken from a corpus of spoken or written language. Since such studies are only beginning to be accumulated for Japanese, it is often the case that an individual researcher cannot find previously compiled word lists, but must conduct the necessary study herself. In particular, these psycholinguistic characteristics are virtually undocumented for Japanese verbs. The present paper is thus a first contribution to this literature, both in the data it provides and, in the absence of any prior model, as a prototype of how to conduct such a quantitative analysis of Japanese verb use.

The need for this corpus study arose in connection with a series of psycholinguistic experiments which studied how Japanese adult native speakers conjugate verbs. Their task was to provide inflected forms[1)] for invented ("nonce") verbs. These "verbs" were created so as to resemble real verbs from three different verb classes ("paradigms"): the two major inflectional verb classes, 1-dan and 5-dan (一段・五段, often called in English "vowel-stem" and "consonant-stem," respectively), plus the irregular verb *suru*. In addition, three different inflections for these verbs appeared in the experiment: the base form (*-ru* or *-suru*), the negative, and the past tense. Therefore, I wanted to know the relative predominance of these inflections and these paradigms, as measured by the number of each found in the language.

There are at least two ways to count language units: by token (延べ語数) and by type (異なり語数). A study of popular magazines (National Language Research Institute, 1964, pp. 66-67),

1) In this paper, "form" or "inflected form" is used to mean a particular pronounceable word (like *yonda*), sometimes also called "token." A group of forms may share a meaning either in the stem (a "lemma," such as *yomu, yonda, yomoo,* etc.) or in a common ending ("inflection," as *yonda, tabeta, shita*).

reported token counts for negative endings and *-ta/-te* endings of 6.5% and 44.4%, respectively, of the 30,701 verb forms in the study (counting only the inflection immediately following the verb stem).[2] The verb type is generally the "lemma," which combines the base form and all the inflected forms which result from it. For instance, the lemma *taberu* includes the forms *tabete, tabeta, tabeyoo, tabetai,* etc. Here are lemmas classified by paradigm, from the same magazine study of contemporary magazines (NLRI, 1964, p. 64), and from a study of classical Japanese texts (Hayashi, 1982, p. 343, citing Keino, 1972):

| *Paradigm* | *Magazines* (n=3,457) | *Classical* (n=13,737) |
|---|---|---|
| 5-dan | 62.9% | 48.3% |
| 1-dan (*-iru*) | 1.9 | 2.7 |
| 1-dan (*-eru*) | 29.7 | 21.1 |
| *suru* | 5.2 | 24.3 |
| Other | 0.2 | 3.5 |
| Total | 99.9% | 99.9% |

For my project, which was a completely aural/oral task, I wanted this kind of information for modern spoken language, and I also wanted information on the frequency of specific inflection forms, like past tense *-ta* vs. *-tta*. So, I undertook to study a small corpus, which will be described in the following section. The corpus was chosen partly because it was a corpus of spoken language which was small and available, and partly because it contained language that a small child heard during his language-learning period. Age of acquisition has been cited as an important influence on language accuracy and fluency throughout life, and it has even been claimed that age of acquisition is a better predictor than is the better-known indicator of frequency (Turner, Valentine & Ellis, 1998). Although these findings were not applied directly to the processing of verb inflections, we can hypothesize that not just individual words, but also other aspects of language which are learned early may retain an advantage over those which are learned later. So I was interested in this child corpus partly as a sample of Japanese in general, and partly as a gauge of language which might be heard and learned in early childhood.

I will describe the corpus, and then give the methodology and rationale that I used to extract, classify, and count the verbs. Finally, the results of these counts will be reported.

## I. The Aki corpus

The "Aki" corpus (Miyata, 1995) is a part of the CHILDES database of child language texts (MacWhinney, 1995). Miyata recorded a Japanese boy in his home more or less weekly, from the age of 1 year, 5 months to 3 years—56 sessions in all. She then transcribed the recordings, using Roman characters, each utterance on a separate line and divided by spaces into wordlike chunks, to make the corpus.

Two aspects of this corpus deserve mention. First, as a corpus of spoken language, it has certain differences from most written texts. Sentences are shorter and less formal than in written language, incorporating exclamations and non-lexical sounds (*a / n / un / aramaa*),[3] and often an utterance does not even contain a complete clause (*nani ga / dame / hai o jii chan*). Secondly, as a corpus of child-directed language, utterances are shorter still, and include some nonstandard words that are specific to child language: *okochiru, nokkatchatta, meme, tete.*

For my purposes, I selected from this corpus only the adult utterances, omitting whatever Aki said, in order to gather a sample of language of the kind that very young children hear and from which they must learn. Adults present at various times were Aki's mother, father, two grandmothers, a grandmotherly baby-sitter, and the researcher. After eliminating 2,364 utterances consisting only of *n* or *un*, the selected utterances numbered 24,073, and 13% of these (3,137) were only a single word (*atta / aramaa / dore,* etc.). There were approximately 77,000 "word" chunks, so on average

2) Actually, what was reported was the occurrence of the stem changes which are required to precede the negative, past, and *-te* endings. The「未然形1A」stem form precedes negative endings *-nai, -zu,* and *-zaru,* etc. The「連用形2」is used for *-te, -ta, -tari,* etc. on verbs, and for *-ta, -tari,* etc. on adjectives.

3) When quoting corpus material, the symbol '/' is used to indicate a line break, which in turn means a new utterance.

there were 3.1 "words" per utterance.

My next task was to extract all the occurrences of verbs. In order to help the reader understand this process, I offer a short explanation of the nature of this corpus. First, here are some typical utterances:[4]

*un takai yo*
*mata machigaechatta*
*jidoosha made ikitai yo tte densha ga yutteru*
*kono hito basu ni noru njanakatta*
*hitori de jakku wa nigemashita*

Although there is no tradition or standard for "word" division in Japanese orthography, the divisions here are made with good consistency. Often morphemes rather than words are separated:[5]

*ree chan tachi* [レーちゃん達]
*booshi no oji san ga mada notte ii no tte kita yo*
*chotto tori nikui ne*
*o baa chan to o jii chan wa oo yorokobi*
*motte kite*
*ki o tsukete*

But in the case of verbs, sequences like the following are represented as single chunks:

*motteiru*
*motteinai*
*ikitakunai*

The transcription of some words seems to reflect variations in pronunciation. Here are the most common variants that were encountered in the Aki corpus.

| | |
|---|---|
| 良い | *ii, yoi* |
| 言う | *iu, yu* |
| 美味しい | *oishii, oishi* |
| 多い | *ooi, ooii* |
| 大きい | *ookii, okii, okkii* |
| 可愛い | *kawaii, kawai* |
| 黄色い | *kiiroi, kiroi* |
| 酸っぱい | *suppai, supai* |
| 小さい | *chiisai, chiitchai, chitchai, chiichai, chitcha na* |
| 遠い | *tooi, toi* |
| 難しい | *muzukashii, mutsukashii* |

One frequent variation in transcription occurs with the contraction of ～てしまう. In general, the Romaji spelling *-cha-* is used to indicate ～てしま～, *-tcha-* is used for ～ってしま～ and *-ja-* for ～でしま～:

| | |
|---|---|
| *tondechatta* | [飛んでいちゃった<br>＝*tobu*＋*iru*＋*shimatta*] |
| *tondetchatta* | [飛んで行っちゃった<br>＝*tobu*＋*iku*＋*shimatta*] |

However, these rules are not always followed in the corpus, perhaps because in actual speech an extra glottal stop may be inserted for emphasis:

*ree chan tte naitchau yo*
*ashi ga hiraitchatta*
*o kaa san netchatta*

The same tendency appears occasionally in the body of a word; for example, one word group with common variations in glottalization is: *kuttsuku, kuttsukeru* (25 tokens) vs. *kutsuku, kutsukeru* (7 tokens). Also, the less formal nature of spoken language is reflected in the occurrence of nonstandard words, such as *tsunageru* (＝*tsunagu*) and *tsukameru* (＝*tsukamu*), and alternate usages: both *taberareru* and *tabereru*; both *sasete* and *sashite*.

## II. Extracting verbs

From this subcorpus of adult utterances, I extracted all instances of verbs, plus some additional forms which share with verbs both form and function, as will be described. Selection was by a combination of methods, sometimes using computer software, and sometimes picking and choosing by hand and eye. First I used the Unix command 'grep' (a pattern-matching program) to extract all chunks (strings of characters with no embedded spaces) with typical verb endings (*-ta, -te, -u,* etc.) or internal bits (*-na-*, etc.), and used the command 'uniq' to combine all instances ("tokens") of the same string ("type") and count them. Then, I checked each of these string types individually, discarding any that were clearly not verbs (*mutsu, fooku, manaita*). Where the string was ambiguous with respect to part of speech (*ike*＝池 or 行け; *mise*＝店 or 見せ; *kaeru*＝蛙 or 変える, 帰る), I went

4) This paper is directed at an audience of English-reading Japanese-speaking scholars, so I will omit English glosses of the Japanese words, instead using Japanese orthography when necessary to disambiguate the Romaji forms.

5) The version of the corpus I used is now several years old. In the current version, some of these examples are differently segmented.

Table 1 Classification by morphological group of the verb and verb-use forms extracted from the Aki corpus, with count of types (lemmas) and tokens

| | Lemmas | Tokens | Lemma % | Token % |
|---|---|---|---|---|
| 1-dan | 149 | 4,205 | 33.6 | 19.6 |
| 5-dan | 192 | 6,688 | 43.3 | 31.1 |
| *-bu, -gu, -ku, -mu, -nu* | 59 | 1,781 | | |
| *-ru, -tu, -u* | 91 | 4,400 | | |
| *-su* | 42 | 507 | | |
| *suru* (1,528), *kuru* (203) | 2 | 1,731 | 0.5 | 8.1 |
| Ambiguous verb forms | 0 | 1,756 | 0.0 | 8.2 |
| Verb subtotal | 343 | 14,380 | 77.4 | 67.0 |
| Adjectives | 83 | 3,204 | 18.7 | 14.9 |
| Copula (*da, desu*, etc.) | 1 | 2,182 | 0.2 | 10.2 |
| *tte* | 1 | 1,197 | 0.2 | 5.6 |
| Other idioms and frozen forms | 15 | 509 | 3.4 | 2.4 |
| Grand total | 443 | 21,472 | 99.9 | 100.1 |

back to the corpus and examined each ambiguous occurrence in its context. This selection process was iterative: often while reading the corpus to check a problem in one group, I noticed some other verbs that had not yet been extracted, so a new 'grep' cycle was designed to catch that group. This process, which in retrospect does not seem very systematic, was ultimately quite a thorough one, so that I think very few relevant words were missed. In all, 21,472 occurrences of verbs and verb-like forms were identified, or a little more than one-fourth of the "chunks" in the corpus.

The first results of this extraction process are shown in Table 1, which shows lemma and token counts categorized by morphological group and subgroup.[6] Because the verbs are the most numerous, and because their forms are more complex, they are subdivided into several groups,[7] including one for forms which were ambiguous between two or three inflectional classes or subclasses, and which were too numerous or too difficult to disambiguate from context (these are listed in Table 2).[8]

In addition to verbs, I decided to collect other word groups if they included forms that were likely to be perceived as verbal by the small child. Thus, words which had both a verbal function (meaning) and a verb-like form were sought out. Certainly the class of words called *-i* adjectives overlaps to a great extent with verb form and function in Japanese. Adjectives are used as predicates in Japanese, just as are verbs, and both are conjugated similarly (particularly in the past tense).[9] Further, two common verb forms are *-i* adjectives morphologically: the negative *-nai* and the desiderative *-tai*. In this corpus, base forms (*-i*) accounted for 87% of all adjective tokens, and the following forms also appeared: *-katta, -ku, -kute, -kunai, -kattara, -sugi-*.

Copular predicates (*da, desu, nai*, etc.) similarly serve verbal functions and are inflected like verbs (*datta, nakattara*, etc.), so these form

6) "Morphological group" refers to a group which share the same forms (verbs, adjectives, etc.). Because of the nature of language, these formal groups often coincide with grammatical "parts of speech."

7) All occurrences of *suru* and its inflections were counted, but any nominals which were used with them were not analyzed for this report.

8) For instance, *iru* can mean either 'to be' or 'to need.' The only way to tell is to read each utterance that uses *iru* (and sometimes the preceding and following utterances) and try to figure out from the context which is meant. (Of course, the child has the same problem, and pretty much the same solution!) In each of these ambiguous cases, there are other forms of the lemmas involved which are not ambiguous and are counted normally, so the ambiguous group is shown as "0 lemmas" in Table 1.

9) The so-called "*na*-adjectives" were not included here, since they are not inflected.

Table 2 Ambiguous verb forms, with instance counts, possible lemmas, and lemma verb inflection class

| Verb form | # Instances | Possible lemmas (verb inflection class and subclass) |
|---|---|---|
| *atta* | 266 | *aru* 有る(5r); *au* 会う(5u) |
| *iru* | 221 | *iru* 居る(1); *iru* 要る(5r) |
| *itt-, itch-* | 459 | *iku* 行く(5k); *iu* 言う(5u); *iru* 要る(5r) |
| *katt-* | 59 | *kau* 買う(5u); *katsu* 勝つ(5t) |
| *ki-* | 700 | *kuru* 来る(irr); *kiru* 着る(1) |
| *shimatt-* | 14 | *shimau* 'to finish'(5u); *shimaru* 閉まる(5r) |
| *yond-* | 37 | *yobu* 呼ぶ(5b); *yomu* 読む(5m) |
| Total | 1,756 | |

Note: Where casual inspection suggested that one possibility was preponderant in the corpus, that one is underscored.

another category. The last two groups cover a variety of idiomatic forms, most of which are frozen forms (see Appendix A for a complete list). The most frequent of these is the form *tte*, which is used mainly in two contexts in this corpus: as a pure quotative, followed by a verb of expression like *iu, oshieru, kaku* (243 instances), and as an idiom meaning 'they say' or some similar attribution, where no following verb supplies this meaning explicitly (954 instances). Because these two uses are phonetically and pragmatically so similar (especially to a small child), I have lumped them all together and, because it occurs so often, I have listed it separately in Table 1.

There are two methodological problems that I would like to mention briefly. First, how to determine when related verbs should be combined into a single lemma? In European languages, lemma groups are generally words that are related by both form and meaning, but in Japanese it is not so simple. For instance, the meanings of *hajimeru* and *hajimaru* are closely related, corresponding to the transitive and intransitive uses of 'to begin' in English, but these are two different lemmas in Japanese, distinguished by different conjugations. For this study, I have counted these, as well as even more closely related derived verbs like *iu* and *iwareru*, as separate lemmas, but I'll defer a fuller discussion to Section IV.

A second problem is when to count the *renyoukei* ("deverbal nominal") form, like *tabe, nomi*. This form occurs as an unbound nominal, as part of a verb compound, and in other formations with varying degrees of boundedness. In this study, I have not been totally consistent when counting these forms, but rather was often influenced by the word-spacing decisions reflected in the corpus, and by the number of instances involved. Thus, the few compound verbs which appeared unspaced, as single chunks, were treated as distinct lemmas (rather than tallying them with either or both of the two component verbs): *mochiageru, hamidasu, hikkurikaesu, oikakeru, kumitateru*. The second part of spaced-out compounds was counted if it had a verbal ending, as shown in the lists below (with occurrence counts, and underlining indicating forms which were counted):

| | | | |
|---|---|---|---|
| 1 | *hari tsukeru* | 2 | *yari naoshi* |
| 1 | *ii kakeru* | 1 | *mimi kaki* |
| 1 | *nori kaeru* | 2 | *iki domari* |
| 2 | *omoi dasu* | 1 | *nakama hazure* |
| 1 | *tobi tsuku* | 1 | *haya oki* |
| 1 | *tobi oriru* | 1 | *tobi komi* |
| 1 | *magari kireru* | 3 | *chikara mochi* [adj.] |

However, 291 of these "nominal" forms were counted in uses like the following:

*ja chanto o suwarishite ne*
*tsukamaeni kita*
*kore yarinikui*
*nigenasai yo*

On the other hand, both parts of syntactically combined verbs[10] were counted (underlined parts are counted; "+" shows that two parts of a single chunk were counted in two different lemmas):

10) The first part of the compound ends in *-te* instead of the *renyoukei* form. We did not analyze these as compounds, however; that would be interesting for a future study.

1 *hotte oite*
8 *itte*+*kimasu*
1 *oite*+*oite*
1 *mattetote* (=*matte*+*oite*)
2 *tondetchatta* (=*tonde*+*itchatta*)

Because I developed policies as I went along, inconsistencies were introduced. In future studies, one could settle these matters in advance. However, there will always be a degree of arbitrariness, since these combining forms show a continuum of closeness, from clearly separate to clearly bound, determined by various linguistic factors, including frequency of use.

In passing, I would like to note that there are advantages and disadvantages to a word-spaced Romaji transcription like this one, compared to a transcription in standard Japanese orthography (kanji/kana). First, it is easier to extract the verb forms mechanically without having to tag or parse[11] the entire corpus first. Japanese orthography is not word-spaced, so a preliminary chunking process would be necessary, which generally involves morphological and/or syntactic parsing. In addition, there are at least two ways to write most words in Japanese orthography (e.g. わかれる, 分かれる, 別れる), so a more complex process would be needed to identify and extract all the verb forms. On the other hand, the purely phonetic nature of the Romaji transcription can make it harder to disambiguate forms which sound alike but have quite different meanings. The forms in Table 2, for instance, are not ambiguous if written with kanji.

## III. Classification of inflected forms

The next step was to classify the verbal forms as to their inflectional group. The morphemes of interest in this study were the negative (*-na-*) and the past (*-ta/-da*),[12] but I also analyzed "*-te*-type" forms (*-te/-de*, *-tari/-dari*, *-tara/-dara*), because their formation is parallel to the past formation in most environments. Thus, the first goal of classification was to count the number of verb forms which belong to these three groups, like *tabenai*, *dekita*, and *yonde*. Table 3 shows all the verb forms broken down by their inflection type; the Infl column shows each subgroup coded as N, P, T, or none. But what about a form like *tabenakatta*? It could be counted either as a negative form (*-na-*) or as a past form (*-katta*). I chose to count such forms multiple times, once for each inflectional part (see section (d) in Table 3). Note that the contracted forms (*-cha*, *-chau*, etc.) are not counted as *-te* forms. It is doubtful that the learner can identify these as *-te* forms initially, but only after considerable experience realizes that they are in free variation with *-te shimau*.

In addition to counting the negative, past, and *-te* inflections, I also wanted to know the relative quantities of the several different endings ("inflectional subgroup") within each of these three inflection groups (*V*+*nai* vs. *V*+*anai*, *V*+*ta* vs. *V*+*tta*, etc.). Assigning each form to the proper inflection subgroup sometimes depends only on the ending (*-nakatta* is always a 5-dan-type form) and sometimes upon the paradigm of the preceding verb stem (*tabenai* is 1-dan because *taberu* is an 1-dan verb, while *nomanai* is 5-dan like *nomu*). These relationships are indicated by a second code in Table 3. For instance, the forms *tabenakatta* and *nomanakatta* would both be coded as Nv, P5, meaning that the negative inflection is formed according to the paradigm of the main verb (1 for *tabena-* and 5 for *nomana-*), but the past tense ending is always *-katta*, regardless of the main verb.

We have now defined four different dimensions of classification for these tokens: morphological class (Table 1), complex verb form (Table 3), and inflectional group and subgroup (coded in Table 3, column 'Infl'). Table 4 shows the intersection of the major groups in Tables 1 and 3. Then, the codes in Table 3 are used to generate counts for each of the three inflection groups, and these are displayed in Tables 5, 6, and 7, and a summary of the total number of occurrences from each of the three inflec-

11) "Tag" means to assign a part of speech (e.g. noun, verb, etc.) to each "word"-chunk (or morphemic part of a chunk). "Parse" means to analyze each utterance to determine its morphological and/or syntactic structure.

12) Throughout the paper, when *-t-* forms are discussed, the parallel *-d-* forms are understood to be included, as *yonde*, *yonda*, *yondari*, *yonderu*, etc.

Table 3 Classification of verb tokens by complex verb form types, with inflectional analysis

| Group | Subgroup label and descriptions | Number of tokens | Infl* | Group total |
|---|---|---|---|---|
| a) | No *-na/ta/te* form in the verb complex: | | | 10,072 |
| a1) | Base form | 7,745 | — | |
| a2) | *-tai, -masu, -masen, -mashoo; -ku* (adjectives) | 543 | — | |
| a3) | *-oo, -yoo* [hortative] | 1,282 | — | |
| a4) | *-e* [imperative] | 31 | — | |
| a5) | *-eba* | 9 | — | |
| a6) | Noun forms (incl. adj+*sugi*, V+*nasai*, V+*ni*, etc.) | 324 | — | |
| a7) | *-cha, -tcha* [=*tewa*] | 4 | — | |
| a8) | *-chau* (92), *-tchau* (37), *-jau* (5) [=*teshimau*] | 134 | — | |
| b) | *-na/ta/te* is first auxiliary (no other aux is coded): | | | 10,396 |
| b1) | *-ta* (incl. adj+*katta*) | 3,331 | Pv | |
| b2) | *-tara, -tari* (incl. adj+*kattara*) | 161 | Tv | |
| b3) | *-te, -temo* (44) (incl. adj+*kute*) | 4,415 | Tv | |
| b4) | *-te(i)ru, -teimasu* (5) | 1,075 | Tv | |
| b5) | *-nai, -naide* (62), *-naku* (61), *-nakucha* (2), *-nakya* (11) | 1,414 | Nv | |
| c) | *-na/ta/te* is second auxiliary (no other aux is coded): | | | 705 |
| c1) | *-mashita* | 175 | Ps | |
| c2) | *-takatta* | 7 | P5 | |
| c3) | V+*takunai* (1), adj+*kunai* (71) | 72 | No | |
| c4) | *-chatta* (193), *-tchatta* (239) [=*teshimatta*] | 443 | P5 | |
| c5) | *-(t)chatte* (8) [=*teshimatte*] | 8 | T5 | |
| d) | Multiple auxiliaries are coded (codes shown in order): | | | 299 |
| d1) | *-nakatta* | 42 | Nv, P5 | |
| d2) | *-te(i)ta* | 88 | Tv, P1 | |
| d3) | *-te(i)te, -te(i)tara* | 35 | Tv, T1 | |
| d4) | *-te(i)nai, -tenakya* (1) | 112 | Tv, N1 | |
| d5) | *-te(i)nakatta* | 2 | Tv, N1, P5 | |
| d6) | *-nakute(mo)* | 20 | Nv, To | |
| | Grand total | | | 21,472 |

*Codes for inflectional analysis:
Inflected form: N=negative; P=past; T=*-te*, etc.
Inflection paradigm: v=from stem verb; 1=1-dan; 5=5-dan; s=*su(ru)*; o=other
Note: Adjectival past inflections (*-katta*) are similar to 5-dan *-ru/-u/-tu* and are thus coded as 5-dan.

Table 4 Classification of verb tokens by complex verb form group and morphological class of main verb

| Group | 1-dan | 5-dan | *suru/kuru* | Ambig | Adj | Copula | Idiom/*tte* | Totals |
|---|---|---|---|---|---|---|---|---|
| a) No *-na/ta/te* | 1,331 | 2,701 | 737 | 240 | 3,012 | 1,559 | 492 | 10,072 |
| b) *-na/ta/te* first | 2,603 | 3,559 | 956 | 1,355 | 121 | 588 | 1,214 | 10,396 |
| c) *-na/ta/te* second | 205 | 262 | 16 | 151 | 71 | — | — | 705 |
| d) Multiple *-na/ta/te* | 66 | 166 | 22 | 10 | — | 35 | — | 299 |
| Chunk totals | 4,205 | 6,688 | 1,731 | 1,756 | 3,204 | 2,182 | 1,706 | 21,472 |
| Morpheme totals* | 2,940 | 4,155 | 1,016 | 1,526 | 192 | 658 | 1,214 | 11,701 |

*Note: Groups b), c), and d), with counts for d) multiplied to reflect multiple morphemes.

tion groups appears as Table 8. Table 8 shows that the most numerous group is the *-te*-type forms (51%), followed by the past (35%), and negatives are by far the fewest (14%), with less than half the number of the past-tense occurrences. If we compute percentages based on the total number of endings in the corpus (21,773), the proportions become 7.6% for negative and 46.1% for the combined *-ta/-te* endings, which are similar to the NLRI (1964) percentages of 6.5% and 44.4%,

respectively.

Tables 5 through 7 show in detail how the complex verb form types in Table 3 were allocated to inflection groups and subgroups. In each case, three major subgroups were identified, corresponding to 1-dan verb inflections, one or more of the 5-dan verb inflections, and *suru* inflections (often shared by the 5-dan *-su* verbs). While the members of each subgroup are not necessarily related to each other, the child does not know this, and may thus hear them as identical forms. Looking at Tables 5 through 7, are there some forms which are more "typical" of their inflection group than others? Or are the several varieties of each inflection equally prevalent? Of the clearly verbal *-nai* forms (Table 5), there is a predominance of the simpler 1-dan forms, with forms like *dekinai* being heard 60% more often than forms like *shiranai*. On the other hand, both *-ta* and *-te* show a predominance of *-tta/-tte* forms over *-ta/-te*. For past-tense forms (Table 6) the glottalized *-tta* ending, which occurs in 5-dan paradigms *-u*, *-ru*, and *-tu* as well as copula and adjectives, is considerably more numerous than *-ta* (35% greater),

Table 5 Classification of negative morphemes by complex verb form and inflection

| | | *-[e/i]na-* | *-ana-* | *-shina-* | Subtotal* | Other | Total |
|---|---|---|---|---|---|---|---|
| b5, d1, d6) | 1-dan and 5-dan | 451 | 359 | — | 810 | — | 810 |
| | *suru* (*shina-*), *kuru* (*kona-*) | — | — | 116 | 116 | 51 | 167 |
| | Copula (*na-*, *jana-*, etc.) | — | — | — | — | 483 | 483 |
| | Idioms | 9 | — | — | 9 | 7 | 16 |
| c3) | Adjectives (*-kunai*) | — | — | — | — | 72 | 72 |
| d4+d5) | *-te(i)na-* | 114 | — | — | 114 | — | 114 |
| Totals | | 574 | 359 | 116 | 1,049 | 613 | 1,662 |
| Percent of total | | 34% | 22% | 7% | 63% | 37% | 100% |
| Percent of subtotal excl. "other" | | 55% | 34% | 11% | 100% | | |

*Notes: Because the "other" group is a fairly high proportion of the total, and because one-third (562=all "other" except *kona-*) are not associated with canonical verbs, but are independent forms (copular) or occur with adjectives or frozen forms (see Table 5), a subtotal and percentages excluding the "other" category are given. The morpheme *-masen* (only 4 instances) is not included in the negative counts, since it is not at all similar to the majority *-nai* forms.

Table 6 Classification of past-tense morphemes by complex verb form and inflection

| | | *-ta* | *-tta* | *-shita* | Other | Total |
|---|---|---|---|---|---|---|
| b1) | *-ta* (1-dan and 5-dan) | 839 | 633 | 26 | 258 | |
| | *shita* (*suru*) | — | — | 263 | — | |
| | Ambiguous verb forms* | 502 | 554 | — | 9 | |
| | Adjectives (*-katta*) | — | 114 | — | — | |
| | Copula (*(n)datta*) | — | 133 | — | — | 3,331 |
| c1) | *-mashita* | — | — | 175 | — | 175 |
| c2, c4) | *-takatta*, *-(t)chatta* | — | 450 | — | — | 450 |
| d2) | *-te(i)ta* | 88 | — | — | — | 88 |
| d1, d5) | *-nakatta*, *-te(i)nakatta* | — | 44 | — | — | 44 |
| Totals | | 1,429 | 1,928 | 464 | 267 | 4,088 |
| Percents of total | | 35% | 47% | 11% | 7% | 100% |

*Note: Includes 502 instances of *kita*, which might be the past form of irregular *kuru*, or 1-dan *kiru*.

with *-shita* forms distinctly in the minority.

The *te*-type forms are the most numerous of all (Table 7). Here the predominance of *-tte* over *-te* (85% greater) is even greater than for the past tense. One confounding factor, however, may be the large contribution to these counts of the idiom *tte* 'it is said,' which accounts for 20% of all *-te*-type forms. It may be thought that *tte* is heard as a frozen form, quite different from those *-te* forms which are connected to another morpheme, so a second set of totals and percentages is computed excluding that form. In this analysis, *-tte* forms are only 8% greater than *-te*.

When the past and *-te*-type forms are considered together (Table 9), on the assumption that their paradigms are parallel, we see that both the 1-dan and 5-dan paradigms are strongly represented, but there is a predominance of *-tta/tte* forms over *-ta/te*, even when we ignore the stand-alone *tte* forms. This contrasts with negative forms, where the 1-dan *-nai* is more common in this corpus than the 5-dan *-anai*.

Token frequency, which we have been discussing so far, is of course the summation of instan-

Table 7 Classification of -te-type morphemes by complex verb form and inflection

| | | *-te* | *-tte* | *-shite* | Other | Total |
|---|---|---|---|---|---|---|
| b2) *-tara, -tari* | | 43 | 69 | 33 | 16 | 161 |
| b3) *-te, -temo:* | 1-dan, 5-dan | 1,128 | 933 | 279 | 323 | |
| | *suru* | — | — | 322 | — | |
| | Ambiguous | 118 | 91 | — | 13 | |
| | *tte*, adj, copula | — | 1,197 | — | 11 | 4,415 |
| b4) *-te(i)ru/masu:* | 1-dan, 5-dan | 168 | 448 | 28 | 205 | |
| | *suru* | — | — | 182 | — | |
| | Ambiguous | 11 | 22 | — | 11 | 1,075 |
| d2-d5) * | All other-*te(i)*- | 58 | 118 | 14 | 47 | 237 |
| c5, d3, d6) * | Third-morpheme *-te*-type | 35 | 8 | — | 20 | 63 |
| Totals with *tte* | | 1,561 | 2,886 | 858 | 646 | 5,951 |
| Percents of total | | 26% | 48% | 14% | 11% | 99% |
| Totals excluding *tte* | | 1,561 | 1,689 | 858 | 646 | 4,754 |
| Percents of total excluding *tte* | | 33% | 36% | 18% | 14% | 101% |

*Note that d3) *-te(i)te* is counted in two categories, since it contains two *-te* morphemes.

Table 8 Summary of negative, past, and *-te*-type morphemes by inflection type

| | 1-dan | 5-dan | *-su/suru* | Other | Total | Percent |
|---|---|---|---|---|---|---|
| | *-[e/i]na-* | *-ana-* | *-shina-* | | | |
| Negative morphemes | 574 | 359 | 116 | 613 | 1,662 | 14% |
| Percent of group total | 34% | 22% | 7% | 37% | 100% | |
| | *-ta* | *-tta* | *-shita* | | | |
| Past-tense morphemes | 1,429 | 1,928 | 464 | 267 | 4,088 | 35% |
| Percent of group total | 35% | 47% | 11% | 7% | 100% | |
| | *-te* | *-tte* | *-shite* | | | |
| *-te*-type morphemes | 1,561 | 2,886 | 858 | 646 | 5,951 | 51% |
| Percent of group total | 26% | 48% | 14% | 11% | 99% | |
| Percent of group excl. *tte* | 33% | 36% | 18% | 14% | 101% | |
| Totals | 3,564 | 5,173 | 1,438 | 1,526 | 11,701 | 100% |
| Percents of total | 31% | 44% | 12% | 13% | 100% | |

Table 9 Summary of past and *-te*-type morphemes (excluding *tte*) by inflection

| | *-ta/te* | *-tta/tte* | *-shita/te* | Other | Total |
|---|---|---|---|---|---|
| Past | 1,429 | 1,928 | 464 | 267 | 4,088 |
| *-te*-type excluding *tte* (1197) | 1,561 | 1,689 | 858 | 646 | 4,754 |
| Totals | 2,990 | 3,617 | 1,322 | 893 | 8,822 |
| Percents of total | 34% | 41% | 15% | 10% | 100% |

Table 10 Exclusion of individual high-frequency types from past and *-te*-type frequency counts

| Form (in order of frequency) | | *-ta/te* | *-tta/tte* | *-shita/te* | Other | Total |
|---|---|---|---|---|---|---|
| Past | *kita* | 502 | — | — | | |
| | *atta* | — | 266 | — | | |
| | *shita* | — | — | 263 | | |
| | *itta* | — | 256 | — | | |
| | *dekita* | 225 | — | — | | |
| Totals | | 727 | 522 | 263 | | 1,512 |
| *-te*-type | *shite* | — | — | 318 | | |
| | *yatte* | — | 315 | | | |
| | *dete* | 239 | — | — | | |
| | *misete* | 204 | — | — | | |
| | *shiteiru* | — | — | 180 | | |
| | *motte* | — | 158 | — | | |
| Totals | | 443 | 473 | 498 | | 1,414 |
| Inclusive totals from Table 9: past | | 1,429 | 1,928 | 464 | 267 | 4,088 |
| *-te*-type (excluding *tte*) | | 1,561 | 1,689 | 858 | 626 | 4,734 |
| Totals excl. high-freq. types: past | | 702 | 1,406 | 201 | 267 | 2,576 |
| *-te*-type (excluding *tte*) | | 1,118 | 1,216 | 360 | 626 | 3,320 |
| New grand totals | | 1,820 | 2,622 | 561 | 893 | 5,896 |
| Percents of total | | 31% | 44% | 10% | 15% | 100% |

ces of wordforms ("types") which are not identical, but which are similar in some important respect.[13] However, Bybee (1995) suggests that the appropriate frequency calculation should exclude those types which are individually most frequent: "...items that are of high frequency in the input can be learned on their own terms, while lower-frequency items are better learned in relation to existing items" (p. 429). To test this hypothesis for Japanese, I made a new calculation of token frequency (Table 10), eliminating those types which are individually most frequent in the corpus. Of the roughly 1500 verb form types in the Aki corpus, I considered as "highly frequent" those types which had a frequency of 145 or greater (1% of the total 14,380 verb tokens). There are 18 such types, of which 11 involve *-te* or *-ta*, 6 are base forms (*-u*), and two are *-yoo* (one form, *shiteiru*, is both *-te* and *-ru*). If we compare Table 10 with Table 9, we see that removing the high-frequency forms from token frequency counts does not alter the dominance of *-tta/tte* forms, but makes them even more predominant over the *-ta/te* forms.

## IV. Lemma frequency and derived verbs

Another way of assessing the "strength" of each paradigm is the number of different verbs that belong to each one. The more distinct verbs that

13) The distinction between "lemma" and "type" is an important one that may be confusing. A "type," as used here, is a unique wordform, while a "lemma" is a collection of a number of such related types. Thus, the lemma *taberu* includes the types *taberu, tabeta, tabenai,* etc.

participate in a common inflection pattern, the more strongly that pattern is thought to be fixed in mind. But what does it mean to be a "distinct verb"? In Table 1, each distinct verb was called a "lemma," and the number of lemmas in each morphological category was given along with the token frequencies. In Section II, I explained how I decided what to include in a lemma group, particularly how I dealt with compounds and deverbal nominals, and here I would like to treat the problem of derived verbs.

In Japanese, inflected forms with passive, causative, and potential meanings can be created for almost any verb, not by adding another morpheme to the standard inflected forms, but by creating a new base verb and then inflecting it. That is, the path to create the derived forms is from the base form to a new derived base (*-ru*) form, and then to inflections:

| Source (1-dan) | | Derived (1-dan) |
|---|---|---|
| *taberu* | → | *tabesaseru* |
| ↓ | | ↓ |
| *tabeta* | | *tabesaseta* |
| *tabereba* | | *tabesasereba* |
| etc. | | etc. |

| Source (5-dan) | | Derived (1-dan) |
|---|---|---|
| *nomu* | → | *nomareru* |
| ↓ | | ↓ |
| *nonda* | | *nomareta* |
| *nomeba* | | *nomarereba* |
| etc. | | etc. |

In our corpus, the following derived verb patterns appear (examples are shown):

| | |
|---|---|
| Passive/potential: | |
| 1-dan/1-dan | *taberu/taberareru* |
| | *neru/nereru/* |
| | *okiru/okireru* |
| 5-dan/1-dan | *hasamu/hasamareru/* |
| | *okoru/okorareru* |
| | *aruku/arukeru,* |
| | *hairu/haireru* |
| Causative:[14] | |
| 1-dan/5-dan | *taberu/tabesashite* |
| 5-dan/5-dan | *tsukuru/tsukurashite* |
| 5-dan/1-dan | *kaku/kakaseru* |

In the canonical case of such a derived verb, both form and meaning are related to the source verb in a wholly predictable fashion, like the examples above, and it could be argued that the source and its derived verbs all belong to the same lemma. Indeed, the NLRI (1964) magazine study treated derived verbs as part of the base verb lemma, with a subsidiary annotation giving their percentage contribution to the total lemma frequency. In my study, since derived verbs often have a different inflectional paradigm from the base, I have generally separated them into their own lemmas.

Over time, a derived verb may acquire more and more of "a life of its own," developing idiosyncratic meanings which augment, or eventually replace, the predictable compositional meaning. Because this process of lexical change is ongoing, it is sometimes difficult to say which verbs should still be classed as derived and which have become independent. In the Aki corpus, I have classed the following as independent verbs rather than as derived: *ikeru, kireru, oreru, toreru, umareru.* In addition, there are cases where a derived and a non-derived verb are homophonous and cannot be distinguished except by context. In the Aki corpus, there were three of these that were each treated as a single, ambiguous lemma: *tateru* (5 tokens: 立てる, 建てる), *kakeru* (70 tokens: 掛ける, 書ける, 駆ける, 欠ける, etc.), and *toreru* (112 tokens: 取れる, 捕れる, 撮れる), while a few others were determined by context to be derived verbs:

| Romaji (#tokens) | Counted as derived | Other possibilities |
|---|---|---|
| *aeta* (1) | 会えた | 敢えた |
| *harenai* (1) | 張れない | 晴れない |
| *moteru* (2) | 持てる | もてる('to be popular') |

Appendix B lists all the verb lemmas in the Aki corpus, arranged by inflection paradigm. All verbs classed as derived are noted there, mainly in two special "derived" sections—sublists of 1-dan *-eru* (43 verbs) and 5-dan *-su* (12 verbs). In all, there are 55 derived verbs listed, accounting for 184 tokens. This is a very tiny proportion of all

14) Martin (1987, p. 287): "Colloquially the suffix [-(s)ase-] is often shortened to -(s)as-, not to be confused with various suffixes that derive transitive verbs and are historically related." In this corpus, there are 32 tokens of causatives, all *-te* form, and 29 of them end in *-sashite.*

the tokens in the corpus, so the reclassification of the derived verbs can make no difference in our analysis of token frequency. However, it does change somewhat the proportion of lemmas in each verb category.

Table 11 shows that, by lemma count, 5-dan verbs are more numerous than the next largest group, the 1-dan *-eru* verbs, while the 1-dan *-iru* verbs are quite few, no matter which corpus or counting method. However, if the derived verbs are not given their own lemmas, but are counted as part of the base verb lemma, the dominance of 5-dan becomes even stronger. This is the method used in the magazine study (NLRI, 1964, p. 64), where the 5-dan verb lemmas are twice the number of both 1-dan groups. When comparing these proportions with our corpus, the only surprise is that the 1-dan *-iru* verbs are an even smaller percentage in the magazine corpus, implying that a higher proportion of these verbs are very common, even in language to young children, than in the other groups. Indeed, the number of 1-dan *-iru* verbs in the Aki corpus amounts to 24% of the number in the magazine corpus, compared to only 8-9% of the other categories.

## V. Conclusions

While studies of frequency of occurrence in Japanese have been published, these have all focused on lemmas rather than inflected forms. There are few statistics available on the frequency of use of particular forms of Japanese verbs and adjectives. This paper has explored some of the theoretical issues and practical problems involved in gathering such statistics.

Overall, what have we learned from this analysis of three types of common verbal inflection? Foremost, we found that the bulk (by token frequency) of these inflections can be categorized into two main groups, resembling the 1-dan and 5-dan paradigms, with a much smaller number of tokens representing other paradigms, including the *-su/suru* group. Of these two main groups, the 1-dan forms predominate in the negative, while the 5-dan forms predominate in the past and *-te*-type groups. Secondly, we saw that the *-te*-type forms are the most frequent of the three inflection groups, accounting for half of all the tokens examined, followed by past tense, with the negative forms much less frequently used. It may also be of interest to note that particular lexical items occur more often in one tense than another, idiosyncratically. For instance, while both *kita* and *kite* appeared in the list of high-frequency forms, *itta* but not *itte* appeared in that list, and *motte* but not *motta*, etc.

Thirdly, we saw from a lemma count of the verbs in this corpus that there are almost twice as many 5-dan verbs represented as 1-dan verbs, and that this proportion is similar to the one reported for an adult corpus of written Japanese.

## References

Bybee, Joan. 1995 Regular morphology and the lexicon. *Language and Cognitive Processes* **10**, 425-455.

Hayashi Ooki. 1982 *Japanese in diagrams: Seeing language graphically.* Tokyo: Kadogawa Shoten. [In Japanese: 林大, 図説日本語—グラフで見る

Table 11 A comparison of lemma counts: derived verbs counted separately vs. combined, and counts from an adult corpus (National Language Research Institute 1964)

| Group | Derived verbs separate | | Derived verbs with base verb | | Magazine corpus | | Percent |
|---|---|---|---|---|---|---|---|
| | # lemmas | % | # lemmas (A) | % | # lemmas (B) | % | A/B |
| 1-dan *-iru* | 15 | 4.4 | 16 | 5.5 | 66 | 2.0 | 24 |
| 1-dan *-eru* | 134 | 39.1 | 91 | 31.3 | 1,028 | 31.4 | 9 |
| 5-dan* | 192 | 56.0 | 182 | 62.5 | 2,174 | 66.5 | 8 |
| *suru* | 1 | 0.3 | 1 | 0.3 | — | — | |
| *kuru* | 1 | 0.3 | 1 | 0.3 | — | — | 8 |
| Totals | 343 | 100.1 | 291 | 99.9 | 3,268 | 99.9 | 9 |

*Note: The NLRI magazine study lists 4-dan rather than 5-dan, and I assume they are equivalent. It also lists 189 verbs of other paradigms, which are not included here.

ことばの姿，東京：角川書店]

Keino Masaji. 1972 *The study of verbs.* Tokyo: Kasama Shoin. [In Japanese: 慶野正次，動詞の研究，東京：笠間書院]

MacWhinney, Brian. 1995 *The CHILDES project: Tools for analyzing talk,* 2nd ed. Erlbaum.

Martin, Samuel E. 1987 *A reference grammar of Japanese.* Rutland, VT: Charles E. Tuttle Company.

Miyata, Susanne. 1995 The Aki corpus: Longitudinal speech data of a Japanese boy aged 1.6-2.12. *Bulletin of Aichi Shukutoku Junior College* **34**, 183-191.

National Language Research Institute. 1964 *Vocabulary and Chinese characters in ninety magazines of today, Volume III: Analysis of the results* (NLRI Report 25). Tokyo, Japan. [In Japanese: 国立国語研究所，現代雑誌九十種の用語用字，第三分冊・分析]

Turner, Judith E., Tim Valentine, and Andrew W. Ellis. 1998 Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition* **26**, 1282-1291.

—Received Sept. 30, 1999—

## Appendix A: Copular and idiomatic forms

| | | |
|---|---|---|
| **Copula:** | | **2,182** |
| *da, datta, dattara, daroo* | 1,140 | |
| *desu, deshita, deshoo* | 296 | |
| *dewa, dewanai* | 2 | |
| *nanda, nandatta, nanjanai* | 21 | |
| *nda, nde, ndatta, ndaroo* | 129 | |
| *nano* [at end of utterance] | 115 | |
| *nai, nakatta, nakute, naku* | 275 | |
| *janai, janakatta, janakute* | 154 | |
| *njanai, njanakatta, njanakute* | 50 | |
| **Idioms (excluding *tte*):** | | **509** |
| *arigatoo* | 111 | |
| *choodai* | 51 | |
| *gochisoo, gochisoosama* | 9 | |
| *gomen* (alone = 28), *gomennasai* (16) | 44 | |
| *goran* (as V-*te goran*) | 109 | |
| *ir(r)as(s)hai, irrasshaimase, itter(r)s(s)hai* | 37 | |
| *kamoshirenai, kamoshiren* | 9 | |
| *kudasai* | 103 | |
| *mottanaide* | 1 | |
| *-nasai* (except *gomen*) | 10 | |
| *ohayoo gozaimasu* | 11 | |
| *omachidoo* | 1 | |
| *onegai* | 2 | |
| *rasshai* (as V-*te rasshai*) | 2 | |
| *sayo(o)nara* | 2 | |
| *shooganai* | 7 | |
| ***tte*** | | **1,197** |

# Appendix B: List of Verbs

Note: + symbol indicates that additional forms are counted as ambiguous forms (see Table 3 in text).

**irregular** 不規則

| | |
|---|---|
| kuru | 203+ |
| suru | 1,524 |

***-iru* 1-dan** 上一段

| | |
|---|---|
| abiru | 1 |
| akiru | 1 |
| dekiru | 345 |
| iru (居る) | 182+ |
| kariru | 2 |
| kiru (着る) | 0+ |
| miru | 611 |
| niru | 40 |
| ochiru | 65 |
| okiru | 22 |
| okochiru | 9 |
| oriru | 35 |
| sugiru | 32 |
| tariru | 5 |
| tsuujiru | 1 |

***-eru* 1-dan** 下一段
(#=derived/non-derived forms combined; see Section IV)

| | |
|---|---|
| afureru | 1 |
| ageru | 152 |
| akeru | 47 |
| akirameru | 1 |
| ateru | 1 |
| atsumeru | 2 |
| azukeru | 1 |
| chigireru | 2 |
| dekakeru | 7 |
| deru | 367 |
| dokeru | 1 |
| haeru | 1 |
| hajimeru | 10 |
| hazureru | 1 |
| hirogeru | 2 |
| hittsukeru | 4 |
| homeru | 1 |
| ireru | 243 |
| kabuseru | 4 |
| kaeru (変える) | 4 |
| kakeru # | 70 |
| kakureru | 13 |
| kangaeru | 3 |
| katameru | 2 |
| katazukeru | 40 |
| kazoeru | 9 |
| kieru | 18 |
| kikoeru | 5 |
| kireru | 7 |
| kiseru | 1 |
| koboreru | 4 |
| kowareru | 64 |
| kumitateru | 1 |
| kureru | 87 |
| kutabireru | 4 |
| kuttsukeru | 10 |
| kuwaeru | 1 |
| kuzureru | 1 |
| machigaeru | 34 |
| mageru | 1 |
| makeru | 5 |
| marumeru | 1 |
| mieru | 165 |
| miseru | 215 |
| mitsukeru | 59 |
| mochiageru | 3 |
| moeru | 1 |
| mukaeru | 2 |
| mukeru | 3 |
| nagareru | 1 |
| nageru | 8 |
| nameru | 1 |
| naraberu | 23 |
| neru | 40 |
| nigeru | 45 |
| noseru | 120 |
| nukeru | 3 |
| nureru | 5 |
| oboeru | 9 |
| oikakeru | 10 |
| oreru | 6 |
| osaeru | 2 |
| oshieru | 50 |
| sageru | 1 |
| shimeru | 23 |
| suberu | 6 |
| suteru | 11 |
| taberu | 235 |
| taeru | 1 |
| taoreru | 13 |
| tasukeru | 13 |
| tateru # | 5 |
| todokeru | 2 |
| tomeru | 3 |
| toreru # | 112 |
| tsubureru | 5 |
| tsugireru | 1 |
| tsukamaeru | 19 |
| tsukameru | 1 |
| tsukareru | 15 |
| tsukeru | 118 |
| tsumeru | 2 |
| tsunageru | 6 |
| tsureru | 16 |
| ukeru | 1 |
| umareru | 4 |
| wakareru | 1 |
| wareru | 8 |
| wasureru | 24 |
| yabureru | 20 |
| yameru | 18 |
| yattsukeru | 9 |

***-eru* 1-dan** 下一段 **derived**
(†=source verb not in corpus)

| | |
|---|---|
| aeru (<au) | 1 |
| akereru | 1 |
| arukeru | 10 |
| damasareru † | 1 |
| daseru | 4 |
| dereru | 1 |
| haireru | 12 |
| hakoberu | 2 |
| hareru | 2 |
| hasamareru | 1 |
| hashireru | 2 |
| ieru | 1 |
| ikeru | 27 |
| iwareru | 2 |
| kaereru (帰れる) | 1 |
| kakaseru | 2 |
| kamareru | 2 |
| keseru | 2 |
| koreru (<kuru) | 4 |
| moraeru | 2 |
| moteru | 2 |
| nagurareru † | 1 |
| nereru | 1 |
| nigereru | 1 |
| nomeru | 1 |
| noreru | 27 |
| okireru | 1 |
| okorareru | 7 |
| oseru | 1 |
| owareru | 1 |
| sasareru | 2 |
| saseru (<suru) | 1 |
| shinjirareru † | 1 |
| subereru | 1 |
| taberareru | 4 |
| tabereru | 4 |
| toberu | 2 |
| tooreru | 7 |
| torareru | 2 |
| tsukaeru | 1 |
| ugokeru | 2 |
| utaeru | 2 |
| yomeru | 5 |

**5-dan *-bu*** 五段バ行

| | |
|---|---|
| asobu | 90 |
| erabu | 1 |
| hakobu | 41 |
| korobu | 6 |
| musubu | 6 |
| narabu | 44 |
| tobu | 54 |
| yobu | 1+ |
| yorokobu | 10 |

**5-dan *-gu*** 五段ガ行

| | |
|---|---|
| isogu | 11 |
| oyogu | 14 |
| tsunagu | 26 |

**5-dan *-ku*** 五段カ行
* =some forms defective

| | |
|---|---|
| aku | 53 |
| aruku | 20 |

| | |
|---|---|
| doku | 54 |
| fuku | 21 |
| haku | 8 |
| hataraku | 1 |
| hiku | 14 |
| hiraku | 2 |
| hittsuku | 12 |
| iku* | 412+ |
| itadaku* | 2 |
| kaku | 289 |
| katamuku | 1 |
| kawaku | 5 |
| kiku | 21 |
| kuttsuku | 22 |
| maku | 6 |
| muku | 16 |
| naku | 53 |
| nozoku | 1 |
| nuku | 1 |
| oku | 94 |
| saku | 10 |
| suku | 21 |
| tataku | 10 |
| todoku | 20 |
| tsuku | 127 |
| ugoku | 26 |
| waku | 2 |

**5-dan *-mu* 五段マ行**

| | |
|---|---|
| fukuramu | 1 |
| fumu | 10 |
| hasamu | 20 |
| hekomu | 3 |
| hikkomu | 1 |
| kamu | 4 |
| nomu | 13 |
| sumu | 13 |
| tanomu | 1 |
| tsumamu | 1 |
| tsumu | 9 |
| tsutsumu | 4 |
| umu | 2 |
| yamu | 1 |
| yasumu | 0 |
| yomu | 59+ |
| yugamu | 1 |

**5-dan *-nu* 五段ナ行**

| | |
|---|---|
| shinu | 10 |

**5-dan *-ru* 五段ラ行**

| | |
|---|---|
| agaru | 17 |
| aru | 397+ |
| ataru | 11 |
| atsumaru | 1 |
| butsukaru | 11 |
| chigiru | 1 |
| depparu | 1 |
| furu | 24 |
| futoru | 1 |
| gambaru | 3 |
| hairu | 333 |
| hajimaru | 26 |
| hakaru | 3 |
| haru | 26 |
| hashiru | 53 |
| hikaru | 8 |
| hikkakaru | 2 |
| hikkurikaeru | 5 |
| hipparu | 21 |
| ijiwaru | 2 |
| iru (要る) | 25+ |
| kaburu | 11 |
| kaeru (帰る) | 26 |
| kajiru | 3 |
| kakaru | 6 |
| kawaru | 9 |
| keru | 1 |
| kezuru | 3 |
| kiru (切る) | 27 |
| komaru | 21 |
| korogaru | 6 |
| kukuru | 2 |
| magaru | 20 |
| mawaru | 25 |
| mazaru | 1 |
| mekuru | 1 |
| mitsukaru | 2 |
| moguru | 5 |
| naoru | 36 |
| naru | 284 |
| nemuru | 1 |
| nigiru | 2 |
| noboru | 36 |
| nokkaru | 1 |
| nokoru | 2 |
| noru | 379 |
| nuru | 10 |
| odoru | 3 |
| okoru | 9 |
| okuru | 1 |
| oru | 1 |
| owaru | 52 |
| sawaru | 25 |
| shiboru | 1 |
| shimaru | 2+ |
| shiru | 48 |
| suwaru | 32 |
| togaru | 1 |
| tomaru | 36 |
| tooru | 42 |
| toru | 105 |
| tsukamaru | 1 |
| tsukuru | 326 |
| tsunagaru | 4 |
| uru | 36 |
| utsuru | 3 |
| wakaru | 52 |
| wataru | 9 |
| yaburu | 6 |
| yaru | 571 |
| yoru | 17 |

**5-dan *-su* 五段サ行**

| | |
|---|---|
| dasu | 54 |
| hamidasu | 6 |
| hazusu | 1 |
| hikkurikaesu | 2 |
| kaesu | 6 |
| kakusu | 12 |
| kasu | 83 |
| kesu | 12 |
| kobosu | 1 |
| korogasu | 1 |
| kowasu | 13 |
| kurasu | 1 |
| mawasu | 8 |
| nakusu | 1 |
| naosu | 34 |
| narasu | 1 |
| nekasu | 1 |
| okosu | 2 |
| orosu | 3 |
| osu | 28 |
| otosu | 5 |
| sagasu | 184 |
| sasu | 2 |
| taosu | 1 |
| tobasu | 3 |
| ugokasu | 1 |
| wakasu | 1 |
| watasu | 9 |
| yokosu | 4 |
| yurusu | 1 |

**5-dan *-su* 五段サ行 derived (all *-te* forms)**

| | |
|---|---|
| asobashite | 2 |
| dokashite | 6 |
| hashirashite | 1 |
| iresashite | 1 |
| motashite | 2 |
| sawarashite | 2 |
| suberashite | 1 |
| suwarashite | 3 |
| tabesashite | 1 |
| tsukurashite | 1 |
| yarashite | 6 |
| sashite (<suru) | 3 |

**5-dan *-tsu* 五段タ行**

| | |
|---|---|
| katsu | 3+ |
| matsu | 103 |
| motsu | 207 |
| tatsu | 20 |
| utsu | 3 |

**5-dan *-u* 五段ア行**

| | |
|---|---|
| arau | 8 |
| au | 7+ |
| chigau | 171 |
| harau | 11 |
| hirou | 4 |
| iu, yu | 329+ |
| kau | 42+ |
| morau | 64 |
| narau | 1 |
| niau | 2 |
| omou | 16 |
| shimau | 80+ |
| surechigau | 5 |
| tetsudau | 10 |
| tsukau | 8 |
| utau | 33 |